Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 46 (2015) 143 – 150

Procedia Computer Science

International Conference on Information and Communication Technologies (ICICT 2014)

XML URL Classification based on their semantic structure orientation for Web Mining Applications

Krishna Murthy. A[a,*], Suresha[b]

[a]*Research Scholar, Dept of Computer Science, University of Mysore, Mysore 570006, India*
[b]*Professor Dept of Computer Science, University of Mysore, Mysore 570006, India*

Abstract

Since decades, several attempts have been made on Web based research particularly based on HTML web pages because of their more availability. But, W3 consortium stated that HTML do not provide a better description of semantic structure of the web page contents, because of its limited pre-defined tags, semi structured data, case sensitivity and so on. To overcome these draw backs, Web developers started to develop Web page(s) on XML, Flash kind of new technologies. It makes a way for new research methods. In this article, we mainly focuses on XML URL classification based on their semantic structure orientation. Experimental results show that proposed method achieves overall accuracy level of 97.36% of classification.

Peer-review under responsibility of organizing committee of the International Conference on Information and Communication Technologies (ICICT 2014)

*Keywords:*XML URL; Classification; Web Mining; Semantic Structure.

1. Introduction

Nowadays, as we all know research on Web is the emerging field. For example, improving the quality of Web by analyzing Usability Test, Web Information Extraction, Browsing Web on Small Screen Devices[10,11,12,13] like mobile,

* Corresponding author. Tel.: +91-8951930338
*E-mail address:*krishnarjun.research@gmail.com

Peer-review under responsibility of organizing committee of the International Conference on Information and Communication Technologies (ICICT 2014)
doi:10.1016/j.procs.2015.02.005

PDA (Personal Digital Assistance) etc., Tracking Product Opinions by analyzing user reviews etc,. In general, we call it as 'Web Mining'. According to analysis targets, Web Mining can be divided into three different types, which are *Web Usage Mining, Web Structure Mining* and *Web Content Mining*.

W3 (World Wide Web) consortium stated that, HTML has a lot of drawbacks such as limited defined tags, not case sensitive, semi-structured and designed for only to display data with limited options. Later to overcome these difficulties few technologies have been introduced such as XML, Flash (with good design options) and so on[2]. Therefore, Web developers started to migrate to develop Web pages on these kinds of emerging Web Technologies to provide a better description of semantic structure of the web page contents. Therefore, these days we can see more web pages on Web which are developed using XML and Flash technologies[3].

There are many research fields which have been opened on these new technologies. We proposed dataset creation technique for XML URLs[4]. After that we analyzed the data set based on XML semantic structure orientation type. Here, we have categorized our dataset into four types 1) Pure XML Web pages 2) RSS XML Web pages 3) HTML Embedded XML Web pages 4) Code Based/Sitemap XML Web pages. Fig. 1. depicts the clear view of XML URL categories. In this article we mainly focus on XML URL classification by proposing a new method based on their semantic orientation for future Web mining applications such as Web page segmentation, Noise Removal, Web page adaptation, Search Engine Optimization (SEO) and so on.

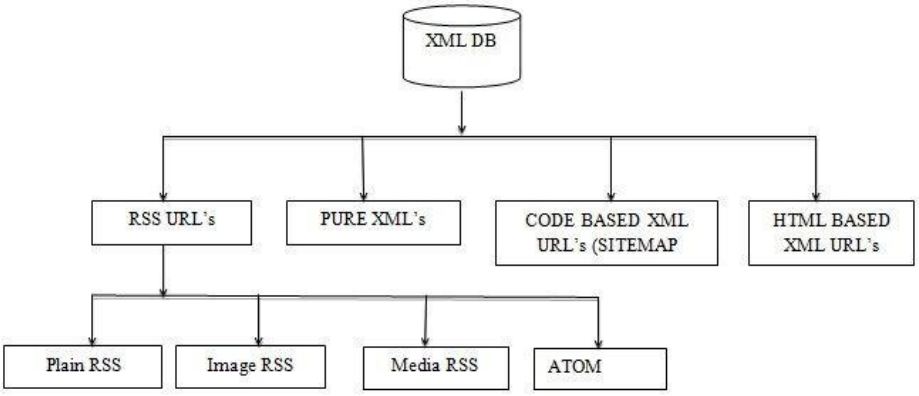

Fig. 1. Dataset Analysis and Classification

Contribution: In light of deficiency of the above mentioned manual process, in this paper we propose an algorithm to Classify the XML URLs based on their semantic structure orientation. Then, we analyze the system accuracy by conducting extensive experiments based on the accuracy measures such as Precision, Recall and F-Measure. Experimental results show that proposed method achieves overall accuracy level of 97.36%.

Organization: After providing the basic information's about XML URLs and its need in research area in Section 1, we present related works in Section 2. We present knowledge base creation method for XML URL classification in Section 3. In Section 4, we describe about training and testing phase of proposed system and in Section 5, we present the result and analysis of conducted experiments on proposed system by using our XML URL Dataset [4].

2. Related Works

In 2003, Vision Based Page Segmentation (VIPS) algorithm[3] proposed to extract the semantic structure of a Web page. Semantic structure is a hierarchical structure in which each node will correspond to a block and each node will be assigned a value to indicate degree of coherence based on visual perception. It may not work well and in many cases the weights of visual separators are inaccurately measured, as it does not take into account the document object model (DOM) tree information and when the blocks are not visibly different.

Gestalt Theory[5]: a psychological theory that can explain human's visual perceptive process. The four basic laws, Proximity, Similarity, Closure and Simplicity are drawn from Gestalt Theory and then implemented in a program to simulate how human understands the layout of Web pages. A graph-theoretic approach[6] is introduced based on

formulating an appropriate optimization problem on weighted graphs, where the weights capture if two nodes in the DOM tree should be placed together. Liu et al.,[7] proposed a novel Web page segmentation algorithm based on finding the Gomory-Hu tree in a planar graph. The algorithm initially distils vision and structure information from a Web page to construct a weighted undirected graph, whose vertices are the leaf nodes of the DOM tree and the edges represent the visible position relationship between vertices. It then partitions the graph with the Gomory-Hu tree based clustering algorithm. Since the graph is a planar graph, the algorithm is very efficient.

From literature[1] it has been observed that no concrete work has been done on Flash Web pages. Hence here we concentrated to work on XML Web page classification for future research avenues.

3. XML URL Classification based on their semantic orientation

System Architecture of proposed system, explains the steps we followed to achieve the classification process as shown in Fig. 2. Each individual process carried out based on XML web pages. Each step is discussed in the upcoming sections.

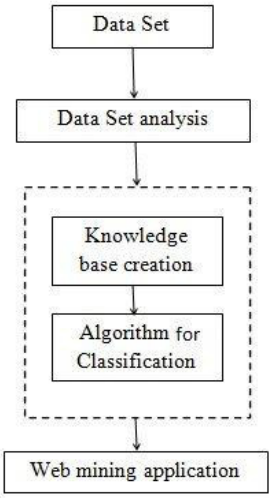

Fig. 2. Architecture of the Proposed System

3.1 Knowledge base

It is a domain knowledge that is used to guide the search or evaluate the interestingness of resulting patterns. Such knowledge can include concept hierarchies, used to organize attributes or attribute values into different levels of abstraction. Knowledge such as user beliefs, which can be used to assess a pattern's interestingness based on its unexpectedness, may also be included. Other examples of domain knowledge are additional interestingness constraints or thresholds, and metadata (e.g., describing data from multiple heterogeneous sources). Here, Knowledge Base is created in four steps as follows.

- Redundancy is checked on XML URL dataset
- Source code extraction
- Tag extraction using DOM structure
- Knowledge Base creation by tag redundancy analysis

3.1.1 XML URL Redundancy Analysis

In our proposed classification method, redundancy analysis is the very first step in Knowledge Base creation task. After creating the various types of XML URL data set such as Pure XML, Code based XML, HTML embedded XML and RSS XML URLs are processed individually in this phase.

Here in first step, Algorithm reads the URL from the source files (Pure XML, Code based XML, HTML embedded XML and RSS XML) line by line and fetch(s) the URL(s). The fetched URL will be tested with destination file for redundancy based on sequential search. If the fetched URL is not present in destination file, then it will be appended otherwise it will not be appended. This process will be continued until the last URL in the source file. Finally the unique XML URLs of each category is obtained.

3.1.2 Source Code Extraction

The resultant vector of first step of the Algorithm will be given as input to the second step of Algorithm to extract the source of respective unique URLs. Here, Algorithm will read the URLs from input file and using Transmission Control Protocol (TCP) it will extract the source code. Extracted source code is saved in auto created destination file with respect to URL number.

3.1.3 Tag extraction using DOM Structure

After extracting the source code of XML URLs, in third step we extract the tags using Document Object Model (DOM) tree structure. Here, the extracted source codes are read line by line and algorithm looks for the tags using DOM. Then, found tags are extracted and stored in corresponding created file name.

3.1.4 Unique Tag Identification

Resultant vector of Step 3 is processed here to identify the unique tags and to create the knowledge base. In this phase, we read each tag files and compare with the destination file tags. Append if the comparing tag does not exist at destination file otherwise skip and move to the next tag of source file. This process will be carried out for all tag files and comparison will be done with destination file.

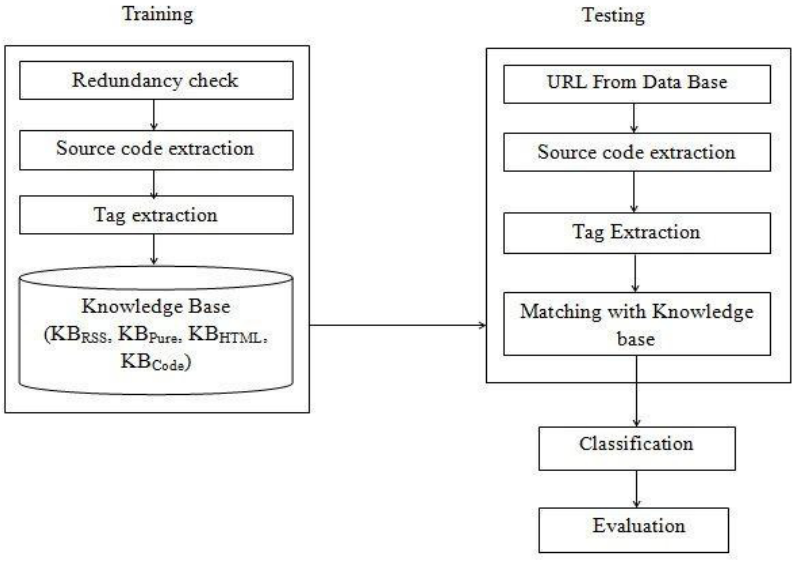

Fig. 3. Block diagram of XML URL Classification

All these four steps are carried out on each type of XML URLs consecutively to create the tag dictionary (Knowledge base). After creating the knowledge base for each category of XML URL's, here matching and representation has been done by using testing dataset. For each testing XML URLs, source code and its tag are extracted.

Here, the extracted tags are matched with Knowledge Base to identify their respective class. Matching process is done with all four Knowledge Bases such as KB_{RSS}, KB_{Pure}, KB_{HTML}, and KB_{Code} by using string matching. Overall matching level is calculated by number tags matched over number of tags of source file. Here the most matched (highest percentage) one is considered as its class.

The above process is carried out for our entire testing data set. Partition of training and testing data sets are given in result analysis section. Fig. 3. depicts the entire structure of Training and Testing data set process with Knowledge Base.

4. Experimentation and Result analysis

In this section, extensive experimentation has been conducted on our proposed model and evaluated the obtained results using accuracy measures such as Precision, Recall and F-measure. For different evaluation purpose our data set has been split into three criteria like 70:30, 60:40 and 50:50 of Training: Testing dataset respectively.

4.1. Accuracy Analysis

Training dataset is the set of data that we use to train the system. It is basically used in various areas of information science. Testing dataset is the set of data used in various areas of information science to check the validation of the system which is trained based on the training dataset. Theoretically, 20% of the data is used for training the system and the rest of the 80% of data is used to test the validation of the system[15]. But, it is not a feasible fact in practical.

Hence, we have considered three categories of data viz., 70:30, 60:40 and 50:50. Where, 70, 60 and 50 refers to the percent of URLs we have considered to train the system and 30, 40 and 50 refer to the percent of URLs that we have used to test the validation of the trained system. The results obtained after training and testing processes is discussed in the following sections.

4.1.1. Results obtained from 70:30 dataset

Table 1. Results obtained for 70:30 dataset.

XML URLs	True Positive	True Negative	False Positive	False Negative	Precision	Recall	F-Measure	Accuracy
CODE	21	286	24	0	0.4600	1.0000	0.6300	92.7
HTML	150	161	0	20	1.0000	0.8800	0.9370	93.9
PURE	4	319	0	8	1.0000	0.3300	0.5000	97.6
RSS	134	197	0	0	1.0000	1.0000	1.0000	100
Avg							**0.7667**	**96.4%**

In the Table 1, 70% of the data is considered as training set and the rest (30%) is used as testing data. With this set of data, we have achieved an average accuracy of **96.4%** and average F-measure of **0.7667**. Graph has been plotted for obtained F-Measure and Accuracy as shown in Fig. 4.

For few category our proposed algorithm achieves less recall and precision value(s) because of tag similarity with other category XML URLs miss classification occurs.

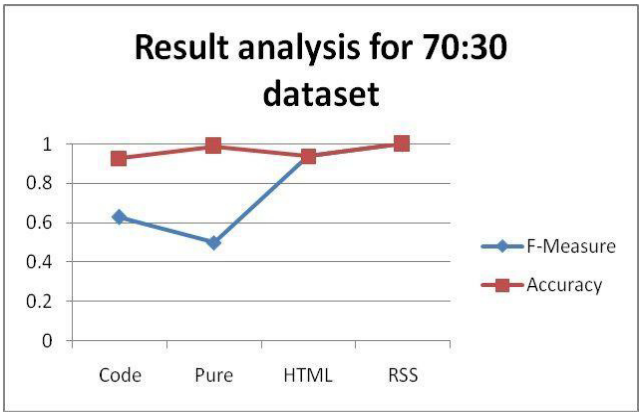

Fig. 4. Accuracy analysis for 70:30.

4.1.2.Results obtained from 60:40 Dataset

Table 2. Results obtained for 60:40 dataset.

XML URLs	True Positive	True negative	False Positive	False Negative	Precision	Recall	F-Measure	Accuracy
CODE	26	378	11	1	0.7021	0.9622	0.8120	97.11
HTML	221	184	0	11	1.0000	0.9521	0.9752	97.35
PURE	8	403	0	5	1.0000	0.6153	0.7611	98.79
RSS	139	261	11	5	0.9261	0.9651	0.9450	96.15
Avg							**0.8731**	**97.35%**

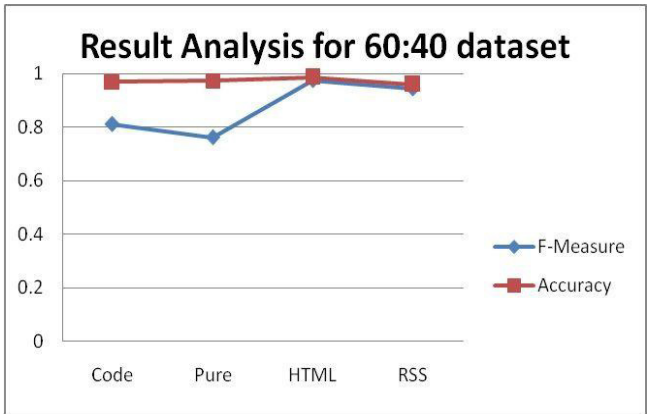

Fig. 5. Accuracy analysis for 60:40.

In Table 2, 60% of the data is considered as training set and the rest 40% is used as testing data. With this set of data we have achieved an average accuracy of 97.35% and an f-measure of 0.8731**.** Corresponding graph has been plotted for obtained F-Measure and Accuracy as shown in Fig. 5.

4.1.3.Results obtained from 50:50 Dataset

In Table 3, we observe that 50% of the data is considered as training set and the rest 50% is used as testing data. With this set of data we have achieved an average accuracy of **97.36%** and an F-measure of **0.8031.**

Table 3. Results obtained for 50:50 dataset.

XML URLs	True Positive	True negative	False Positive	False Negative	Precision	Recall	F-Measure	Accuracy
CODE	29	485	11	4	0.7250	0.8780	0.7940	97.16
HTML	290	228	0	11	1.0000	0.9630	0.9810	97.91
PURE	6	510	2	11	0.7500	0.3520	0.4812	97.50
RSS	184	329	11	5	0.9430	0.9730	0.9580	96.91
Avg							**0.8031**	**97.36%**

Corresponding graph has been plotted for obtained F-Measure and Accuracy as show in Fig. 6.

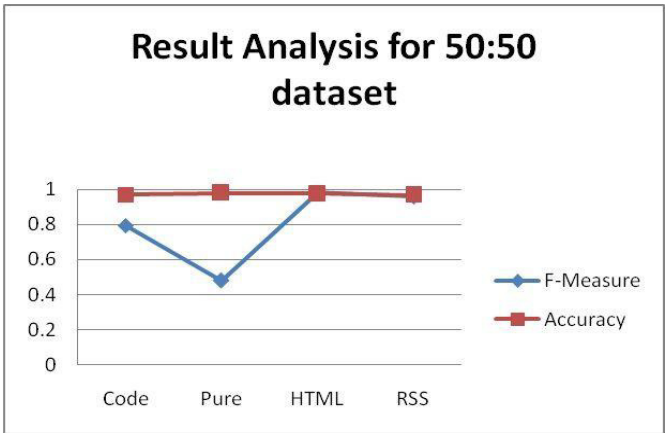

Fig. 6. Accuracy analysis for 50:50.

Table 4. Comparative analysis of Accuracy Levels.

	Accuracy		
Data Set	**70:30**	**60:40**	**50:50**
CODE	92.7	97.11	97.16
HTML	93.9	97.35	97.91
PURE	97.6	98.79	97.50
RSS	100	96.15	96.91
Avg	**96.4%**	**97.35%**	**97.36%**

Finally, we have compared (Table 4) the obtained Accuracy based on various range consideration of data set (70:30, 60:40, 50:50). And we obtained the graph which has been shown in Fig. 7.

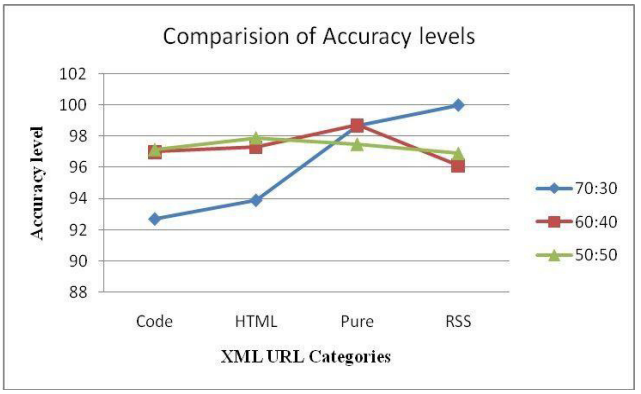

Fig. 7. Comparison of Accuracy levels of all three dataset ration.

5. Conclusion

We have presented a brief overview of importance of classification and its advantages. To achieve the classification system, we proposed four successive Algorithms to create knowledge base. After performing all the four consecutive Algorithms on testing data set elements, matching has been done. Based on the highest matching score Web pages are classified into their respective classes. After proposing an Algorithm, we have conducted extensive experimentation on various ratio of data set and compared the obtained F-Measure and Accuracy score with each other. Overall we have achieved average accuracy of **96.99%** classification with very less error rate.

References

1. Krishna Murthy A, Suresha. Comparative Study on Browsing on Small Screen Devices. *International Journal of Machine Intelligence* ISSN: 0975–2927 and E-ISSN: 0975–9166, 354-358, 2011.
2. Book: Ed Tittel. Complete Coverage of XML, *Tata McGraw*-Hill Edition
3. Deng Cai, Shipeng Yu, Ji-Rong Wen and Wei-Ying Ma.VIPS: A Vision based page segmentation Algorithm. Technical Report MSR-TR-2003-79, 2003.
4. Krishna Murthy. A, Suresha. XML URL Dataset Creation for future Web Mining Research Avenues. *International Journal of Computer Applications* (0975-8887), 2011.
5. Xiang P.F., Yang X. and Shi Y.C. Web page Segmentation based on Gestalt Theory.Conference on *Multimedia and Expo*, 2253-2256 IEEE, 2007.
6. D. Chakrabarti, R. Kumar and K. Punera. A Graph-Theoretic Approach to Webpage Segmentation. *17th International Conference on WWW,* 2008.
7. Xinyue Liu, Xianchao Zhang, Ye Tian. Webpage Segmentation based on Gomory-Hu Tree Clustering. Undirected Planar Graph. *NSFC*, 2010.
8. S. Audithan, RM Chandrasekaran. Document Text Extraction from Document Images using Haar Discrete Wavelet Transform. *European Journal of Scientific Research* ISSN 1450-216X, 2009.
9. Milos Kovacevic, Dilligenti. Recognition of common areas in a web page using visual information: a possible application in a page classification. In *Second IEEE International Conference on Data Mining (ICDM-02),* 2002.
10. P.F Xiang. Effective Page Segmentation Combining Pattern Analysis and Visual Separators for Browsing on Small Screens. *Web Intelligenc*, 2006.
11. Shumeet Baluja. Browsing on small screens: recasting web-page segmentation into an efficient machine learning framework. In *WWW-06: Proceedings of the 15th international conference on World Wide Web*, New York, USA, ACM, 2006.
12. Y. Chen, X. Xie, W.-Y. Ma, and H.-J. Zhang. Adapting web pages for small-screen devices. Internet Computing, 9(1):50–56, 2005.
13. Xin Yang, Yuanchun Shi. Enhanced Gestalt Theory Guided Web Page Segmentation for Mobile Browsing. *IEEE/WIC/ACM,* 2009.
14. Lan Yi, Bing Liu, and Xiaoli Li. Eliminating noisy information in web pages for data mining. In *KDD-03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining,* NY, USA. ACM, 2003.
15. Xavier Amatriain, Alejandro Jaimes and Nuria Oliver. Data Mining Methods for Recommender Systems. Recommender Systems Handbook, *Springer LLC* 2011.